Possible transmission flow of SARS-CoV-2 based on ACE2 features

Sk. Sarif Hassan[a,*], Shinjini Ghosh[b], Diksha Attrish[c], Pabitra Pal Choudhury[d], Vladimir N. Uversky[e,*], Bruce D. Uhal[f], Kenneth Lundstrom[g,*], Nima Rezaei[h], Alaa A. A. Aljabali[i], Murat Seyran[j], Damiano Pizzol[k], Parise Adadi[l], Antonio Soares[m], Tarek Mohamed Abd El-Aziz[n], Ramesh Kandimalla[o], Murtaza Tambuwala[p], Gajendra Kumar Azad[q], Samendra P. Sherchan[r], Wagner Baetas-da-Cruz[s], Kazuo Takayama[t], Ángel Serrano-Aroca[u], Gaurav Chauhan[v], Giorgio Palu[w], Adam M. Brufsky[x]

[a]Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram 721140, India
[b]Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, Kolkata 700009, West Bengal, India
[c]Dr. B. R. Ambedkar Centre For Biomedical Research (ACBR), University of Delhi (North Campus), Delhi 110007, India
[d]Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India
[e]Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[f]Department of Physiology, Michigan State University, East Lansing, MI 48824, USA
[g]PanTherapeutics, Rte de Lavaux 49, CH1095 Lutry, Switzerland
[h]Research Center for Immunodeficiencies, Pediatrics Center of Excellence, Children's Medical Center, Tehran University of Medical Sciences, Tehran, Iran &
Network of Immunity in Infection, Malignancy and Autoimmunity (NIIMA), Universal Scientific Education and Research Network (USERN), Stockholm, Sweden
[i]Department of Pharmaceutics and Pharmaceutical Technology, Yarmouk University-Faculty of Pharmacy, Irbid 566, Jordan
[j]Doctoral studies in natural and technical sciences (SPL 44), University of Vienna, Austria
[k]Italian Agency for Development Cooperation - Khartoum, Sudan Street 33, Al Amarat, Sudan
[l]Department of Food Science, University of Otago, Dunedin 9054, New Zealand
[m]Department of Cellular and Integrative Physiology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr, San Antonio, TX 78229-3900, USA
[n]Department of Cellular and Integrative Physiology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr, San Antonio, TX 78229-3900, USA &
Zoology Department, Faculty of Science, Minia University, El-Minia 61519, Egypt
[o]CSIR-Indian Institute of Chemical Technology Uppal Road, Tarnaka, Hyderabad-500007, Telangana State, India
[p]School of Pharmacy and Pharmaceutical Science, Ulster University, Coleraine BT52 1SA, Northern Ireland, UK
[q]Department of Zoology, Patna University, Patna-800005, Bihar, India
[r]Department of Environmental Health Sciences, Tulane University, New Orleans, LA, 70112, USA
[s]Translational Laboratory in Molecular Physiology, Centre for Experimental Surgery, College of Medicine, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
[t]Center for iPS Cell Research and Application, Kyoto University, Japan
[u]Biomaterials and Bioengineering Lab, Translational Research Centre San Alberto Magno, Catholic University of Valencia San Vicente Mártir, c/Guillem de Castro 94, 46001 Valencia, Spain
[v]School of Engineering and Sciences, Tecnologico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, 64849 Monterrey, Nuevo León, Mexico
[w]Department of Molecular Medicine, University of Padova, Via Gabelli 63, 35121, Padova, Italy
[x]University of Pittsburgh School of Medicine, Department of Medicine, Division of Hematology/Oncology, UPMC Hillman Cancer Center, Pittsburgh, PA, USA

Abstract

Angiotensin-converting enzyme 2 (ACE2) is the cellular receptor for the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) that is engendering the severe coronavirus disease 2019 (COVID-19) pandemic. The spike (S) protein receptor-binding domain (RBD) of SARS-CoV-2 binds to the three sub-domains viz. amino acids (aa) 22-42, aa 79-84, and aa 330-393 of ACE2 on human cells to initiate entry. It was reported earlier that the receptor utilization capacity of ACE2 proteins from different species, such as cats, chimpanzees, dogs, and cattle, are different. A comprehensive analysis of ACE2 receptors of nineteen species was carried out in this study, and the findings propose a possible SARS-CoV-2 transmission flow across these nineteen species.

*Corresponding author

Email addresses: sarimif@gmail.com (Sk. Sarif Hassan), shinjinighosh2014@gmail.com (Shinjini Ghosh), dikshaattrish@gmail.com (Diksha Attrish), pabitrapalchoudhury@gmail.com (Pabitra Pal Choudhury), vuversky@usf.edu (Vladimir N. Uversky), bduhal@gmail.com (Bruce D. Uhal), lundstromkenneth@gmail.com (Kenneth Lundstrom), rezaei_nima@tums.ac.ir (Nima Rezaei), alaaj@yu.edu.jo (Alaa A. A. Aljabali), muratseyran@gmail.com (Murat Seyran), damianopizzol8@gmail.com (Damiano Pizzol), parise.adadi@postgrad.otago.ac.nz (Parise Adadi), soaresa@uthscsa.edu (Antonio Soares), mohamedt1@uthscsa.edu (Tarek Mohamed Abd El-Aziz), ramesh.kandimalla@iict.res.in (Ramesh Kandimalla), m.tambuwala@ulster.ac.uk (Murtaza Tambuwala), gkazad@patnauniversity.ac.in (Gajendra Kumar Azad), sshercha@tulane.edu (Samendra P. Sherchan), wagner.baetas@gmail.com (Wagner Baetas-da-Cruz), kazuo.takayama@cira.kyoto-u.ac.jp (Kazuo Takayama), angel.serrano@ucv.es (Ángel Serrano-Aroca), gchauhan@tec.mx (Gaurav Chauhan), giorgio.palu@unipd.it (Giorgio Palu), brufskyam@upmc.edu (Adam M. Brufsky)

---

1. Introduction

We have been acquainted with the term beta-coronavirus for about two decades when we first encountered the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) outbreak that emerged in 2002, infecting about 8000 people with a 10% mortality rate [1]. It was followed by the emergence of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in 2012 with 2300 cases and mortality rate of 35% [2]. The third outbreak, caused by SARS-CoV-2, was first reported in December 2019 in China, Wuhan province, which rapidly took the form of a pandemic [3, 4]. To date, this new human coronavirus has affected 36 million people worldwide and is held accountable for over one million deaths [5]. SARS-CoV-2 is an enveloped single-stranded plus sense RNA virus whose genome is about 30kb in length, and which encodes for 16 non-structural proteins, four structural and six accessory proteins [6]. The four major structural proteins which play a vital role in viral pathogenesis are Spike protein (S), Nucleocapsid protein (N), Membrane protein (M), and Envelope protein (E) [7, 8]. SARS-CoV-2 infection is mainly characterized by pneumonia [9]; however, multi-organ failure involving myocardial infarction, hepatic, and renal damage is also reported in patients infected with this virus [10]. SARS-CoV-2 binds to the Angiotensin-converting enzyme 2 (ACE2) receptor on the host cell surface via its S protein [11, 12]. ACE2 plays an essential role in viral attachment and entry [13]. The study of the interaction of ACE2 and S protein is of utmost importance [14, 15].

The S1 subunit of the S protein has two domains, the C-terminal and the N-terminal domains, which fold independently, and either of the domains can act as Receptor Binding Domain (RBD) for the interaction and binding to the ACE2 receptor widely expressed on the surface of many cell types of the host [16, 17]. The human ACE2 protein is 805 amino acid long, containing two functional domains: the extracellular N-terminal claw-like peptidase M2 domain and the C-terminal transmembrane collectrin domain with a cytosolic tail [18]. The RBD of the S protein binds to three different regions of ACE2, which are located at amino acids (aa) 24-42, 79-84, and 330-393 positions present in the claw-like peptidase domain of ACE2 [13]. ACE2 modulates angiotensin activities, which promote aldosterone release and increase blood pressure and inflammation, thus causing damage to blood vessel linings and various types of tissue injury [19]. ACE2 converts Angiotensin II to other molecules and reduces this effect [20]. However, when SARS-CoV-2 binds to ACE2, the function of ACE2 is inhibited and in turn leads to endocytosis of the virus particle into the host cell [21].

Zoonotic transmission of this virus from bat to human and random mutations acquired by SARS-CoV-2 during human to human transmission has also empowered this virus with the ability to undergo interspecies transmission and recently many cases have been reported stating that different species can be infected by this virus [22].

In this study, we aim to determine the susceptibility of other species, whether they bear the capability of being a possible host of SARS-CoV-2. We chose nineteen different species and analyzed the ACE2 protein sequence in relation to the human ACE2 sequence and determined the degree of variability by which the sequences differed from each other. We performed a comprehensive bioinformatics analysis in addition to the phylogenetic analysis based on full-length sequence homology, polarity along with individual domain sequence homology and secondary structure prediction of these protein sequences. These finding could be emerged to six distinct clusters of nineteen species based on the collective analysis and thereby provided a prediction of the interspecies SARS-CoV-2 transmission.

2. Data and Methods

2.1. Data acquisition and findings

The ACE2 protein receptor sequences from nineteen species *Homo sapiens* (Human), *Capra hircus* (Domestic goat), *Pan troglodytes* (Chimpanzee), *Equus caballus* (Horse), *Salmo salar* (Atlantic salmon), *Mesocricetus auratus* (Golden hamster), *Rhinolophus ferrumequinum* (Greater horseshoe bat), *Pteropus alecto* (Black flying fox), *Mustela putorius furo* (Domestic ferret), *Danio rerio* (Zebrafish), *Manis javanica* (Sunda pangolin), *Sus scrofa* (Domestic pig), *Macaca mulatta* (Rhesus macaque), *Bos taurus* (Aurochs), *Pelodiscus sinensis* (Chinese soft-shelled turtle), *Pteropus vampyrus* (Large flying fox), *Rattus norvegicus* (Brown rat), *Felis catus* (Domestic cat), and *Gallus gallus* (Red jungle fowl) were derived from the NCBI database [23]. Nineteen species and their respective ACE2 protein accession IDs with length are presented in Table-1.

Table 1: Nineteen species and their associated ACE2 sequences

Name	Species	ACE2 Accession ID	Length
S1	*Homo sapiens*	NP_001358344.1	805
S2	*Pan troglodytes*	PNI38577.1	805
S3	*Macaca mulatta*	XP_028697658.1	805
S4	*Equus caballus*	XP_001490241.1	805
S5	*Felis catus*	NP_001034545.1	805
S6	*Mesocricetus auratus*	XP_005074266.1	805
S7	*Manis javanica*	XP_017505752.1	805
S8	*Pteropus alecto*	XP_006911709.1	805
S9	*Capra hircus*	AHI85757.1	804
S10	*Sus scrofa*	NP_001116542.1	805
S11	*Mustela putorius furo*	XP_004758943.1	805
S12	*Bos taurus*	NP_001019673.2	804
S13	*Rattus norvegicus*	NP_001012006.1	805
S14	*Pteropus vampyrus*	XP_011361275.1	804
S15	*Rhinolophus ferrumequinum*	XP_032963186.1	805
S16	*Gallus gallus*	XP_416822.2	808
S17	*Pelodiscus sinensis*	XP_006122891.1	808
S18	*Danio rerio*	XP_005169417.1	807
S19	*Salmo salar*	XP_014062928.1	695

The nearest neighborhood phylogeny of the nineteen species derived from the NCBI public server based on ACE2 protein sequence similarity is shown in Fig.1 (A) [24].

Figure 1: ACE2 full-length sequence-based phylogeny among nineteen species (A) and its derived clusters (B)

3

ACE2 sequence similarity among the species derives six clusters as shown in (Fig.1 (B)). The contact residues of the receptor-binding domain (RBD) of the spike protein (YP_009724390.1) of SARS-CoV-2 with the homo sapiens ACE2 interface are presented in Table-2 [13].

Table 2: Contact residues of RBD spike protein of SARS-CoV-2 and Homo sapiens ACE2

SARS-CoV-2 RBD	ACE2 (Homo sapiens)	SARS-CoV-2 RBD	ACE2 (Homo sapiens)
K417	Q24	Q493	Q42
G446	T27	G496	L79
Y449	F28	Q498	M82
Y453	D30	T500	Y83
L455	K31	N501	N330
F456	H34	G502	K353
A475	E35	Y505	G354
F486	E37		D355
N487	D38		R357
Y489	Y41		R393

The three designated domains, D1 (aa 24-42), D2 (aa 79-84) and D3(aa 330-393) respectively contain the residues which bind to the RBD of the S protein.

2.2. Methods

Examining amino acid substitutions: For human ACE2 receptor, substitutions were examined for all species and only those substitutions are accounted for, which occurred in the binding residues in the mentioned three domains D1, D2 and D3 [13]. Based character of the substitutions which interfered with the binding residues of the ACE2 across various species, two types were defined: substitutions affected transmission (M1), and substitutions did not affect transmission (M2).

Multiple sequence alignments and associated phylogenetic trees were developed using the NCBI web-suite across all the individual binding domains D1, D2, and D3 of ACE2 of the nineteen species [25, 26].

K-means clustering: Algorithmic clustering technique derives homogeneous subclasses within the data such that data points in each cluster are as similar as possible according to a widely used distance measure viz. Euclidean distance. One of the most commonly used simple clustering techniques is the *K-means clustering* [27, 28]. The algorithm is described below in brief:

Algorithm: K-means algorithm is an iterative algorithm that tries to form equivalence classes from the feature vectors into K (pre-defined) clusters where each data point belongs to only one cluster [27].

- Assign the number of desired clusters (K) (in the present study, $K = 6$).

- Find centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

- Keep iterating until there is no change to the centroids.

- Find the sum of the squared distance between data points and all centroids.

- Assign each data point to the closest cluster (centroid).

- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

In this present study, nineteen species were clustered using *Matlab* by inputting the distance matrix derived from the feature vectors associated with the three domains of ACE2 across all species.

Secondary structure predictions*:* The secondary structure of full-length ACE2 sequence of all species were predicted using the web-server CFSSP (Chou & Fasman Secondary Structure Prediction Server) [29]. This server predicts secondary structure regions from the protein sequence such as alpha-helix, beta-sheet, and turns from the amino acid sequence [29]. On obtaining the full-length ACE2 secondary structures, individual domains D1, D2, and D3 were cropped for each species.

Bioinformatics features*:* Several bioinformatics features viz. Shannon entropy, instability index, aliphatic index, charged residues, half-life, melting temperature, N-terminal of the sequence, molecular weight, extinction coefficient, net charge at pH7 and isoelectric point of D1, D2, and D3 domains of ACE2 for all nineteen species were determined using the web-servers *Pfeature and ProtParam* [30, 31].

Shannon entropy: It measures the amount of complexity in a primary sequence of ACE2. It was determined using the web-server *Pfeature* by the formula

$$SE = -\sum_{i=1}^{20} p_i log_2(p_i)$$

where p_i denotes the frequency probability of a given amino acid in the sequence [30].

Instability index is determined using the web-server *ProtParam* and it estimates the stability of a protein in a test tube. A protein whose instability index is smaller than 40 is predicted as stable. A value above 40 predicts that the protein may be unstable [30].

Aliphatic index of a protein is defined as the relative volume gathered by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for increasing the thermostability of globular proteins, such as ACE2 [30].

N-terminal: It was reported that the N-terminal of a protein is responsible for its function. For each domain sequence, N-terminal residue was determined using the *Pfeature* [30].

In vivo half-life: The half-life predicts the time it takes for half of the protein amount to degrade after its synthesis in the cell. The N-end rule originated from the observations that the identity of the N-terminal residue of a protein plays an essential role in determining its stability in vivo [32].

Extinction coefficients: The extinction coefficient measures how much light a protein absorbs at a particular wavelength. It is useful to estimate this coefficient when a protein is getting purified [32].

Polarity sequence*:* Every amino acid in the domains D1, D2 and D3 of ACE2 were recognized as polar (P) and non-polar (Q) and thus every D1, D2 and D3 for all the species turned out to be binary sequences with two symbols P and Q. Then homology of these sequences for each domain was made and consequently, phylogenetic relationship was drawn.

3. Results

Based on amino acid homology, secondary structures, bioinformatics and polarity of the three domains D1, D2 and D3 of ACE2, all nineteen species were clustered. Finally a cumulative set of nineteen species clusters was built, among which the SARS-CoV-2 transmission may occur.

3.1. Phylogeny and clustering based on ACE2 domain-based homology

First, we examined all the substitutions with similar properties and similar side chain binding atoms, signifying that the substitutions would not impede the SARS-CoV-2 transmission. Note that, all the mutations are considered concerning the human ACE2 domains D1, D2 and D3 (Fig.2).

5

Sl. No.	SPECIES	D1 (24-42 aa)	D2 (79-84 aa)	D3 (330-393 aa)
1	Pan troglodytes	0	0	0
2	Macaca mulatta	0	0	0
3	Equus caballus	Q24L, D30L, H34S, D38E, Y41H	M82T	0
4	Felis catus	Q24L, D30E, D38E	M82T	0
5	Mesocricetus auratus	H34Q	M82T	0
6	Manis javanica	Q24E, D30E, H34S, D38E	L79I, M82N	G354H
7	Pteropus alecto	Q24L, D30E, H34T	M82A	N330K, R393K
8	Capra hircus	D30E	L79M, M82T	0
9	Sus scrofa	Q24L, D30E, H34L	L79I, M82T	R393K
10	Mustela putorius furo	Q24L, D30E, H34Y, D38E	L79H, M82T	G354R
11	Bos taurus	D30E	L79M, M82T	0
12	Rattus norvegicus	Q24K, T27S, D30N, H34Q	L79I, M82N, Y83F	K353H
13	Pteropus vampyrus	Q24L, D30E, H34T	M82A	N330K, R393K
14	Rhinolophus ferrumequinum	Q24L, T27K, K31D, H34S, D38N, Y41H	M82N, Y83F	0
15	Gallus gallus	DEL24Q, D30A, K31E, H34V, E35R, Q42E	L79N, M82R, Y83F	G354N
16	Pelodiscus sinensis	Q24E, T27N,D30S, K31E, H34V, E35Q, Q42A	L79N, M82K	G354K
17	Danio rerio	Q24R, T27E, D30N, H34E, E37S	L79E, M82A	INSERT353N, K353R, G354K
18	Salmo salar	Domain does not exist	Domain does not exist	N330D, INSERT353N, K353R, G354E

Figure 2: Substitutions in D1, D2 and D3 domains of ACE2 across eighteen species

In D1 domain: out of eighteen species, eight species were found to possess a substitution at position 30 where D (Aspartate) was substituted by E (Glutamate) and four species were found to carry the D38E substitution. It was reported that in the aspartate side chain, the oxygen atom was involved in ionic-ionic interaction and the side-chain oxygen atom was also present in glutamate, so this substitution may not affect the protein-protein interaction properties [33, 34]. In the T27S substitution, Threonine and Serine both possess OH that participates in binding and in the H34L substitution, both Histidine and Leucine use the NH group for interaction with another amino acid. Consequently, if we consider only the critical perspective for these substitutions, we can conclude that these changes would not impede the binding between the S and ACE2 protein.

In D2 domain: L79I bears importance across eighteen species since both these amino acids (Leucine and isoleucine) share similar chemical properties. So, if we analyze the changes in amino acid residues based on their chemical properties, which is the main contributing factor for protein-protein interaction, we can conclude that it will not significantly affect the binding between ACE-2 and RBD of the S protein.

In D3 domain: out of eleven substitutions, three substitutions (R393K, K353H, and K353R) were observed of the similar type with similar side chain interacting atoms and therefore changes at these positions would not affect the interaction of ACE2 with that of the S protein.

Secondly, across all nineteen species, homology were derived based on amino acid sequences, and consequently, associated phylogenetic trees were drawn (Fig.3).
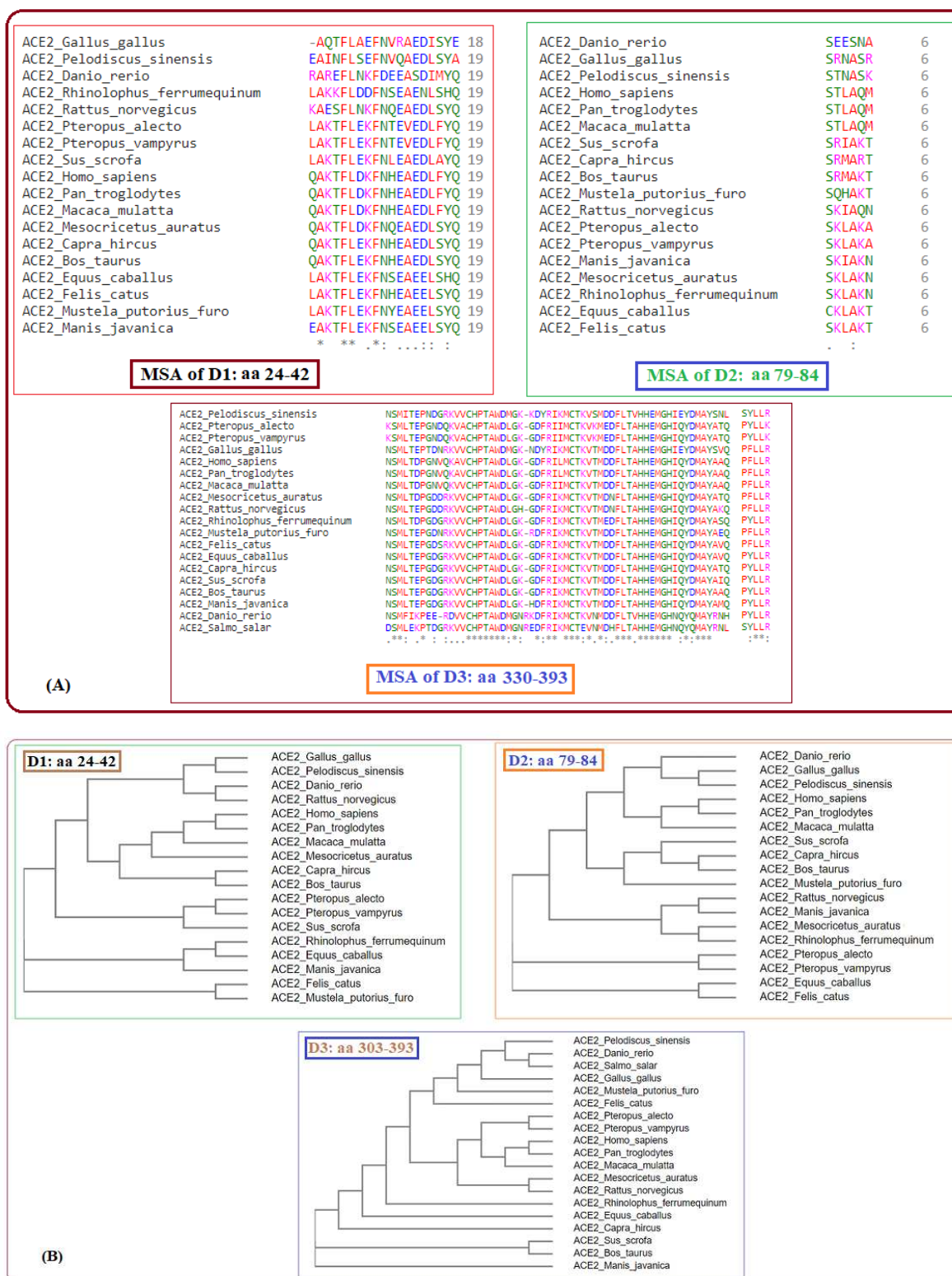
Figure 3: Multiple sequence alignments of D1, D2 and D3 domains of ACE2 of nineteen species (A) and respective phylogenies (B).

Six clusters of the nineteen species were formed using the K-means clustering technique based on sequence homology of the three domains (Fig 4). The clusters of species $\{S1, S2, S3\}$ and $\{S6, S13\}$ stayed together for the ACE2 full-length sequence homology and the combination of three domain-based sequence similarity. The species S16, S17, and S18 also followed the same as observed.
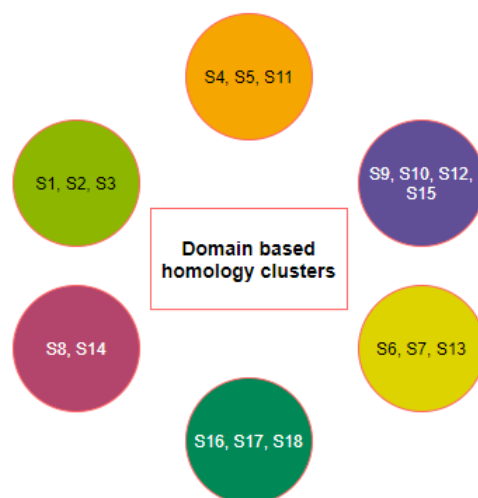
Figure 4: Clusters of species based on domain-based sequence homology

Further, it was observed that sequence homology of the D1, D2, and D3 domains clustered the species S15 into the cluster where S9, S10 and S12 belong, although S15 was similar to the ACE2 sequence of S8 and S9. Although the species S4 was very much similar to S9, S10, and S12 for full-length ACE2 homology, it clubbed with S5 and S11 concerning the three domain-based sequence spatial organizations. Also S7 was found to be staying in the proximity of S6 and S13 although S7 was very much similar to S5 and S11 based on ACE2 homology.

3.2. Clustering based on secondary structures

For each domain of ACE2 of nineteen species, the secondary structure was predicted (Fig.5). For each domain species are grouped into several subgroups.

With respect to the D1 domain:

- *Bos taurus* and *Capra hircus*

- *Equus caballus* and *Felis catus*

- *Mustela putorius furo* has a structure closer to *Equus caballus* and *Felis catus* as it has only one difference of a coil present at position 42. Also, *Mesocricetus auratus* has a secondary structure similar to the above two except an extended helix at position four. Similarly, *Sus scrofa* which has an extended sheet instead of a helix at position 42. So, *Mustela putorius furo*, *Mesocricetus auratus*, and *Sus scrofa* can be put in the same cluster as *Equus caballus* and *Felis catus*.

- *Homo sapiens, Macaca mulatta* and *Pan troglodytes*

- *Manis javanica* and *Rhinolophus ferrumequinum*

- *Pteropus alecto* and *Pteropus Vampyrus*

- *Rattus norvegicus* and *Pelodiscus sinensis* have similar structures differing by the presence of an extra coil at position 39 for *Rattus norvegicus.*

- *Gallus gallus* and *Danio rerio* has a unique secondary structure in comparison to the others.

8

These individual six clusters show six different secondary structures in D1 shared by sixteen species, which shows high similarities in their secondary structures, while the rest two have a unique secondary structure for D1 domain. Thus, these six clusters have similar secondary structures indicating that the species in the six clusters are closely related.

Figure 5: Predicted secondary structures of D1, D2, and D3 domains for all the species

With respect to the D2 domain:

- *Homo sapiens, Macaca mulatta* and *Pan troglodytes*

- *Bos taurus, Mustela putorius furo, Pteropus alecto* and *Pteropus vampyrus*

- *Equus caballus, Felis catus, Manis javanica, Pelodiscus sinensis* and *Rhinolophus ferrumequinum*

155 - *Danio rerio* and *Gallus gallus*

Similarly, for D2 domain, we found four clusters with the same secondary structure, indicating that they are closely related.

With respect to the D3 domain:

- *Homo sapiens* and *Pan troglodytes*

160 - *Bos Taurus, Rhinolophus ferrumequinum, Sus scrofa* and *Capra hircus*

- *Equus caballus* and *Felis catus*

- *Pteropus alecto* and *Pteropus vampyrus*

Again for D3 domain, four different clusters were bearing similar secondary structures therefore these species are also closely related.

165 Based on the similarity among the three domains, all nineteen species were clustered (Fig.6).
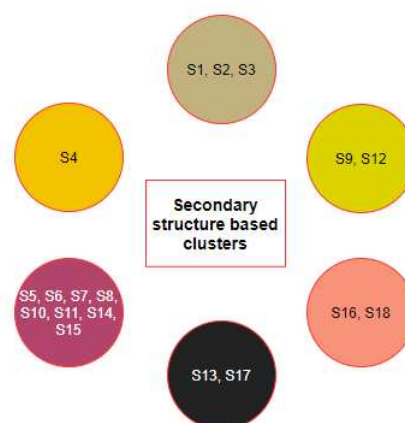
Figure 6: Clusters of species based on the secondary structures of the D1, D2 and D3 domains

From the clusters (Fig.6) based on the secondary structure of the three domains of ACE2, it was observed that the species S4 was clustered uniquely, though S4 is clustered with S9 and S19 based on ACE2 full-length sequence homology. Furthermore, S6 and S13 were found to be similar based on ACE2 homology, but they got clustered into two different clusters when the secondary structure of three domains was concerned. In contrast, the group of species $\{S1, S2, S3\}$, $\{S9, S12\}$, 170 $\{S16, S18\}$, and $\{S5, S7S11\}$ remained in the same clusters concerning to ACE2 homology as well as individual secondary structures of the domains.

3.3. Clustering based on bioinformatics

Twelve bioinformatics features viz. Shannon entropy, Instability Index, aliphatic index, charged residues, half-life, melting temperature, N-terminal of the sequence, molecular weight, extinction coefficient, net charge at pH7, and isoelectric point of the D1, D2, and D3 domains of ACE2 for all nineteen species were determined (Fig.7).

Sl No	SPECIES	Domain	Shannon Entropy	Instability index:	Aliphatic index	Negatively charged residues	Positively charged residues	N-terminal of the sequence	The estimated half-life	Melting Temperature Prediction	MW (g/mol)	Extinction coefficient E(M-1cm-1)	Net Charge at pH7	Estimated Solubility in Water	pI
1	Homo sapiens	D1	3.366	38.79 (Stable)	51.58	4	2	Q	0.8 hours	55℃~65℃	2344.53	1280	-1.9	Good	4.42
2	Pan troglodytes	D1	3.366	38.79 (Stable)	51.58	4	2	Q	0.8 hours	55℃~65℃	2344.53	1280	-1.9	Good	4.42
3	Macaca mulatta	D1	3.366	38.79 (Stable)	51.58	4	2	Q	0.8 hours	55℃~65℃	2344.53	1280	-1.9	Good	4.42
4	Equus caballus	D1	3.156	43.26 (Unstable)	72.11	4	2	L	5.5 hours	lower than 55℃	2221.42	0	-1.9	Good	0.61
5	Felis catus	D1	3.261	33.13 (Stable)	72.11	4	2	L	5.5 hours	55℃~65℃	2297.52	1280	-1.9	Good	3.93
6	Mesocricetus auratus	D1	3.366	27.81 (Stable)	51.58	4	2	Q	0.8 hours	higher than 65℃	2275.43	1280	-2	Good	3.93
7	Manis javanica	D1	3.111	43.26 (Unstable)	51.58	5	2	E	1 hour	lower than 55℃	2263.41	1280	-3	Good	3.97
8	Pteropus alecto	D1	3.287	48.93 (Unstable)	82.11	4	2	L	5.5 hours	higher than 65℃	2335.6	1280	-2	Good	4.04
9	Capra hircus	D1	3.471	26.11 (Stable)	51.58	4	2	Q	0.8 hours	higher than 65℃	2298.46	1280	-1.9	Good	4.52
10	Sus scrofa	D1	3.116	26.11 (Stable)	97.89	4	2	L	5.5 hours	higher than 65℃	2243.5	1280	-2	Good	4.04
11	Mustela putorius furo	D1	3.156	29.16 (Stable)	72.11	4	2	L	5.5 hours	55℃~65℃	2323.55	2560	-2	Good	4.15
12	Bos taurus	D1	3.471	26.11 (Stable)	51.58	4	2	Q	0.8 hours	higher than 65℃	2298.46	1280	-1.9	Good	4.52
13	Rattus norvegicus	D1	3.261	48.38 (Unstable)	51.58	4	2	K	1.3 hours	lower than 55℃	2261.4	1280	-2	Good	4.04
14	Pteropus vampyrus	D1	3.287	48.93 (Unstable)	82.11	4	2	L	5.5 hours	higher than 65℃	2335.6	1280	-2	Good	4.04
15	Rhinolophus ferrumequinum	D1	3.261	15.64 (Stable)	72.11	4	2	L	5.5 hours	lower than 55℃	2206.37	0	-2	Good	4.42
16	Gallus gallus	D1	3.53	22.81 (Stable)	76.11	4	1	A	4.4 hours	lower than 55℃	2103.24	1280	-3	Good	3.54
17	Pelodiscus sinensis	D1	3.326	34.3 (Stable)	92.63	4	0	E	1 hour	lower than 55℃	2160.29	1280	-4	Good	0.53
18	Danio rerio	D1	3.577	58.05 (Unstable)	51.58	5	3	R	1 hour	lower than 55℃	2362.57	1280	-2	Good	4.16
1	Homo sapiens	D2	2.585	8.33 (Stable)	81.67	0	0	S	1.9 hours	lower than 55℃	649.76	0	0	Poor	3.44
2	Pan troglodytes	D2	2.585	8.33 (Stable)	81.67	0	0	S	1.9 hours	lower than 55℃	649.76	0	0	Poor	3.44
3	Macaca mulatta	D2	2.585	8.33 (Stable)	81.67	0	0	S	1.9 hours	lower than 55℃	649.76	0	0	Poor	3.44
4	Equus caballus	D2	2.252	-5.82 (Stable)	81.67	0	2	C	1.2 hours	lower than 55℃	662.84	0	1.9	Good	10.23
5	Felis catus	D2	2.252	-5.82 (Stable)	81.67	0	2	S	1.9 hours	lower than 55℃	646.78	0	2	Good	10.69
6	Mesocricetus auratus	D2	2.252	-5.82 (Stable)	81.67	0	2	S	1.9 hours	lower than 55℃	659.78	0	2	Good	10.69
7	Manis javanica	D2	2.252	-5.82 (Stable)	81.67	0	2	S	1.9 hours	higher than 65℃	659.78	0	2	Good	10.69
8	Pteropus alecto	D2	1.918	-5.82 (Stable)	98.33	0	2	S	1.9 hours	lower than 55℃	616.75	0	2	Good	12.1
9	Capra hircus	D2	2.252	61 (Unstable)	16.67	0	2	S	1.9 hours	lower than 55℃	720.85	0	2	Good	11.39
10	Sus scrofa	D2	2.585	40.43 (Unstable)	81.67	0	2	S	1.9 hours	lower than 55℃	674.79	0	2	Good	11.39
11	Mustela putorius furo	D2	2.585	40.43 (Unstable)	16.67	0	1	S	1.9 hours	higher than 65℃	670.72	0	1.1	Good	9.86
12	Bos taurus	D2	2.585	61 (Unstable)	16.67	0	2	S	1.9 hours	higher than 65℃	692.83	0	2	Good	11.39
13	Rattus norvegicus	D2	2.585	-5.82 (Stable)	81.67	0	1	S	1.9 hours	lower than 55℃	659.73	0	1	Good	9.86
14	Pteropus vampyrus	D2	1.918	-5.82 (Stable)	98.33	0	2	S	1.9 hours	lower than 55℃	616.75	0	2	Good	10.69
15	Rhinolophus ferrumequinum	D2	2.252	-5.82 (Stable)	81.67	0	2	S	1.9 hours	lower than 55℃	659.78	0	2	Good	10.69
16	Gallus gallus	D2	1.918	93.1 (Unstable)	16.67	0	2	S	1.9 hours	lower than 55℃	689.72	0	2	Good	12.1
17	Pelodiscus sinensis	D2	2.252	-16.72 (Stable)	16.67	0	1	S	1.9 hours	lower than 55℃	606.63	0	1	Good	9.86
18	Danio rerio	D2	1.918	126.87 (Unstable)	16.67	2	0	S	1.9 hours	lower than 55℃	635.58	0	-2	Good	0.76
1	Homo sapiens	D3	4.132	36.32 (Stable)	71.72	7	5	N	1.4 hours	55℃~65℃	7253.32	8250	-1.8	Poor	6.05
2	Pan troglodytes	D3	4.132	36.32 (Stable)	71.72	7	5	N	1.4 hours	55℃~65℃	7253.32	8250	-1.8	Poor	6.05
3	Macaca mulatta	D3	4.161	31.98 (Stable)	74.69	7	5	N	1.4 hours	higher than 65℃	7281.37	8250	-1.8	Poor	6.05
4	Equus caballus	D3	4.153	40.09 (Unstable)	67.03	8	7	N	1.4 hours	55℃~65℃	7341.43	9530	-0.8	Good	6.5
5	Felis catus	D3	4.178	49.11 (Unstable)	67.03	8	7	N	1.4 hours	higher than 65℃	7355.46	8250	-0.8	Good	6.5
6	Mesocricetus auratus	D3	4.134	33.55 (Stable)	62.5	8	7	N	1.4 hours	55℃~65℃	7370.43	8250	-0.8	Good	6.5
7	Manis javanica	D3	4.143	40.23 (Unstable)	62.5	8	7	N	1.4 hours	55℃~65℃	7453.59	9530	-0.7	Good	6.61
8	Pteropus alecto	D3	4.135	45.32 (Unstable)	65.62	8	7	K	1.3 hours	55℃~65℃	7356.48	9530	-0.7	Good	6.5
9	Capra hircus	D3	4.143	38.91 (Stable)	62.5	8	7	N	1.4 hours	55℃~65℃	7343.4	9530	-0.8	Good	6.5
10	Sus scrofa	D3	4.161	40.09 (Unstable)	68.59	8	7	N	1.4 hours	55℃~65℃	7355.46	9530	-0.8	Good	6.5
11	Mustela putorius furo	D3	4.186	52.52 (Unstable)	62.5	9	8	N	1.4 hours	55℃~65℃	7511.6	8250	-0.8	Good	6.5
12	Bos taurus	D3	4.147	40.09 (Unstable)	64.6	8	7	N	1.4 hours	55℃~65℃	7313.38	9530	-0.8	Good	6.5
13	Rattus norvegicus	D3	4.172	45.48 (Unstable)	62.5	8	7	N	1.4 hours	55℃~65℃	7420.49	8250	-0.7	Good	6.61
14	Pteropus vampyrus	D3	4.135	45.32 (Unstable)	65.62	8	7	K	1.3 hours	55℃~65℃	7356.48	9530	-0.7	Good	6.5
15	Rhinolophus ferrumequinum	D3	4.172	40.09 (Unstable)	62.5	8	7	N	1.4 hours	55℃~65℃	7329.38	9530	-0.8	Good	6.5
16	Gallus gallus	D3	4.171	47.3 (Unstable)	59.38	9	7	N	1.4 hours	55℃~65℃	7534.61	9530	-1.8	Good	6.06
17	Pelodiscus sinensis	D3	4.119	43.94 (Unstable)	63.91	9	8	N	1.4 hours	55℃~65℃	7524.62	10810	-0.8	Good	6.5
18	Danio rerio	D3	4.18	28.67 (Stable)	51.72	8	9	N	1.4 hours	higher than 65℃	7797.92	9530	8.02	Good	1.3
19	Salmo salar	D3	4.18	29.68 (Stable)	54	9	8	D	1.1 hours	higher than 65℃	7770.81	9530	-0.6	Good	6.62

Figure 7: Bioinformatics of the D1, D2, and D3 domains of ACE2 from nineteen species

For each species, a twelve-dimensional feature vector was found (Fig.7). For each domain D1, D2, and D3 domain, distance matrix was determined using the Euclidean distance

$$d(S,T) = \sqrt{\sum_{i=1}^{12}(f_i - g_i)^2}$$

. Note that here f_i and g_i denote the ith feature for the species S and T respectively. This distance matrices with heatmap representation for all three domains are presented in Figs. 8, 9 & 10. In addition, by inputting the distance matrix, using the K-means clustering technique, several clusters of species were formed for each domain (Figs. 8, 9 and 10).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	0	0	1.286	0.05182	0.06997	0.08126	0.03372	0.04778	0.112	1.28	0.04778	0.08368	0.03372	1.288	0.2431	0.1889	0.02643
2	0	0	0	1.286	0.05182	0.06997	0.08126	0.03372	0.04778	0.112	1.28	0.04778	0.08368	0.03372	1.288	0.2431	0.1889	0.02643
3	0	0	0	1.286	0.05182	0.06997	0.08126	0.03372	0.04778	0.112	1.28	0.04778	0.08368	0.03372	1.288	0.2431	0.1889	0.02643
4	1.286	1.286	1.286	0	1.282	1.281	1.281	1.285	1.283	1.281	2.562	1.283	1.281	1.285	0.03168	1.286	1.282	1.288
5	0.05182	0.05182	0.05182	1.282	0	0.03098	0.04135	0.04242	0.02223	0.06027	1.28	0.02223	0.04446	0.04242	1.283	0.1946	0.1389	0.07278
6	0.06997	0.06997	0.06997	1.281	0.03098	0	0.01963	0.07086	0.0231	0.05647	1.281	0.0231	0.0249	0.07086	1.282	0.174	0.1225	0.09225
7	0.08126	0.08126	0.08126	1.281	0.04135	0.01963	0	0.07873	0.03906	0.05345	1.282	0.03906	0.00569	0.07873	1.282	0.1634	0.1114	0.1003
8	0.03372	0.03372	0.03372	1.285	0.04242	0.07086	0.07873	0	0.05343	0.09619	1.28	0.05343	0.08035	0	1.287	0.2339	0.1763	0.04201
9	0.04778	0.04778	0.04778	1.283	0.02223	0.0231	0.03906	0.05343	0	0.07203	1.28	0	0.04324	0.05343	1.284	0.1968	0.1445	0.07164
10	0.112	0.112	0.112	1.281	0.06027	0.05647	0.05345	0.09619	0.07203	0	1.283	0.07203	0.05458	0.09619	1.281	0.142	0.08402	0.1318
11	1.28	1.28	1.28	2.562	1.28	1.281	1.282	1.28	1.28	1.283	0	1.28	1.282	1.28	2.563	1.299	1.291	1.281
12	0.04778	0.04778	0.04778	1.283	0.02223	0.0231	0.03906	0.05343	0	0.07203	1.28	0	0.04324	0.05343	1.284	0.1968	0.1445	0.07164
13	0.08368	0.08368	0.08368	1.281	0.04446	0.0249	0.00569	0.08035	0.04324	0.05458	1.282	0.04324	0	0.08035	1.282	0.1621	0.1101	0.1016
14	0.03372	0.03372	0.03372	1.285	0.04242	0.07086	0.07873	0	0.05343	0.09619	1.28	0.05343	0.08035	0	1.287	0.2339	0.1763	0.04201
15	1.288	1.288	1.288	0.03168	1.283	1.282	1.282	1.287	1.284	1.281	2.563	1.284	1.282	1.287	0	1.284	1.281	1.29
16	0.2431	0.2431	0.2431	1.286	0.1946	0.174	0.1634	0.2339	0.1968	0.142	1.299	0.1968	0.1621	0.2339	1.284	0	0.06068	0.2629
17	0.1889	0.1889	0.1889	1.282	0.1389	0.1225	0.1114	0.1763	0.1445	0.08402	1.291	0.1445	0.1101	0.1763	1.281	0.06068	0	0.2078
18	0.02643	0.02643	0.02643	1.288	0.07278	0.09225	0.1003	0.04201	0.07164	0.1318	1.281	0.07164	0.1016	0.04201	1.29	0.2629	0.2078	0

Distance matrix derived from bioinformatics feature vectors of D1

Clusters based on bioinformatics of D1

Figure 8: Distance matrix based on the bioinformatics feature vectors of D1 of ACE2 across eighteen species and associated clusters

13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	0	0	0.02063	0.01642	0.01901	0.01901	0.04035	0.1102	0.04157	0.07575	0.09447	0.01852	0.04035	0.01901	0.1144	0.0822	0.136
2	0	0	0	0.02063	0.01642	0.01901	0.01901	0.04035	0.1102	0.04157	0.07575	0.09447	0.01852	0.04035	0.01901	0.1144	0.0822	0.136
3	0	0	0	0.02063	0.01642	0.01901	0.01901	0.04035	0.1102	0.04157	0.07575	0.09447	0.01852	0.04035	0.01901	0.1144	0.0822	0.136
4	0.02063	0.02063	0.02063	0	0.01608	0.003174	0.003174	0.04902	0.1098	0.04779	0.08018	0.09794	0.003496	0.04902	0.003174	0.1214	0.08664	0.1506
5	0.01642	0.01642	0.01642	0.01608	0	0.013	0.013	0.03434	0.1191	0.05408	0.08331	0.104	0.01306	0.03434	0.013	0.1259	0.07719	0.1486
6	0.01901	0.01901	0.01901	0.003174	0.013	0	0	0.04614	0.1115	0.04863	0.08054	0.09891	0.001674	0.04614	0	0.1221	0.08468	0.1501
7	0.01901	0.01901	0.01901	0.003174	0.013	0	0	0.04614	0.1115	0.04863	0.08054	0.09891	0.001674	0.04614	0	0.1221	0.08468	0.1501
8	0.04035	0.04035	0.04035	0.04902	0.03434	0.04614	0.04614	0	0.1482	0.07607	0.1083	0.1301	0.04613	0	0.04614	0.1476	0.08302	0.1573
9	0.1102	0.1102	0.1102	0.1098	0.1191	0.1115	0.1115	0.1482	0	0.08228	0.05425	0.02803	0.1115	0.1482	0.1115	0.04472	0.1382	0.1085
10	0.04157	0.04157	0.04157	0.04779	0.05408	0.04863	0.04863	0.07607	0.08228	0	0.06516	0.07052	0.04868	0.07607	0.04863	0.08499	0.1102	0.1156
11	0.07575	0.07575	0.07575	0.08018	0.08331	0.08054	0.08054	0.1083	0.05425	0.06516	0	0.03027	0.08053	0.1083	0.08054	0.05606	0.08587	0.09383
12	0.09447	0.09447	0.09447	0.09794	0.104	0.09891	0.09891	0.1301	0.02803	0.07052	0.03027	0	0.09894	0.1301	0.09891	0.03227	0.1161	0.08806
13	0.01852	0.01852	0.01852	0.003496	0.01306	0.001674	0.001674	0.04613	0.1115	0.04868	0.08053	0.09894	0	0.04613	0.001674	0.1221	0.08464	0.15
14	0.04035	0.04035	0.04035	0.04902	0.03434	0.04614	0.04614	0	0.1482	0.07607	0.1083	0.1301	0.04613	0	0.04614	0.1476	0.08302	0.1573
15	0.01901	0.01901	0.01901	0.003174	0.013	0	0	0.04614	0.1115	0.04863	0.08054	0.09891	0.001674	0.04614	0	0.1221	0.08468	0.1501
16	0.1144	0.1144	0.1144	0.1214	0.1259	0.1221	0.1221	0.1476	0.04472	0.08499	0.05606	0.03227	0.1221	0.1476	0.1221	0	0.1377	0.06499
17	0.0822	0.0822	0.0822	0.08664	0.07719	0.08468	0.08468	0.08302	0.1382	0.1102	0.08587	0.1161	0.08464	0.08302	0.08468	0.1377	0	0.1468
18	0.136	0.136	0.136	0.1506	0.1486	0.1501	0.1501	0.1573	0.1085	0.1156	0.09383	0.08806	0.15	0.1573	0.1501	0.06499	0.1468	0

Distance matrix derived from bioinformatics feature vectors of D2

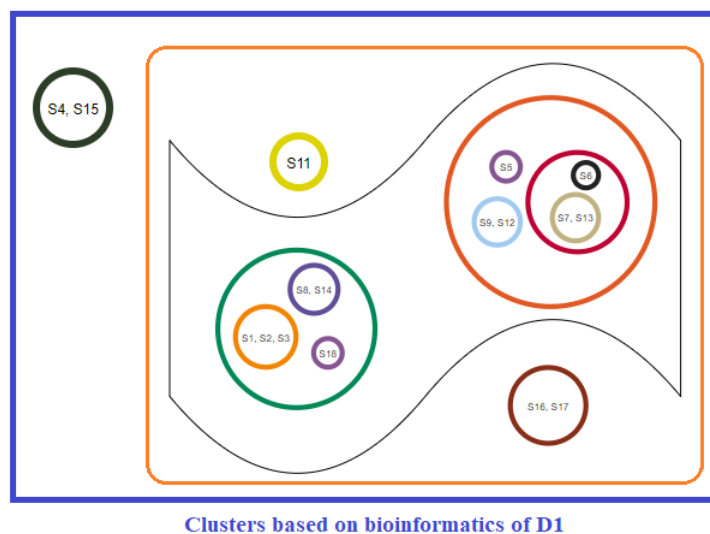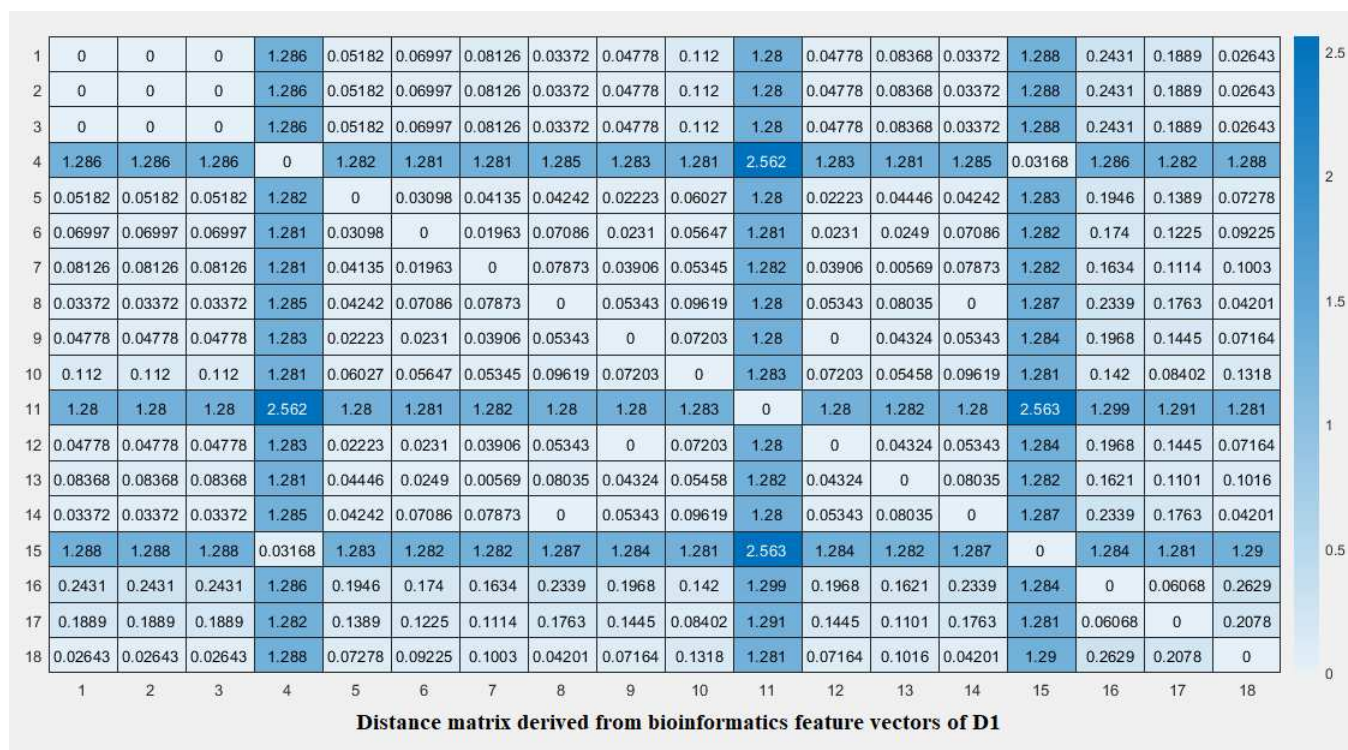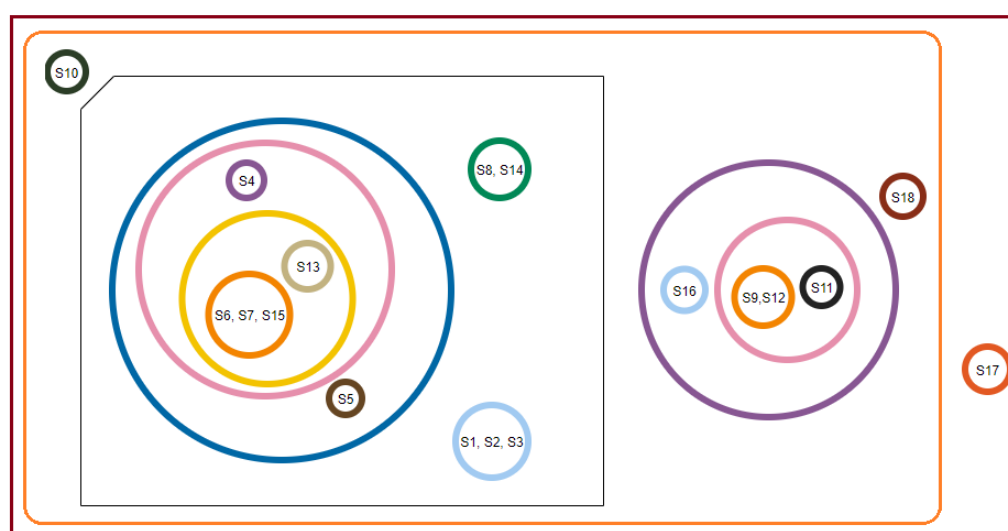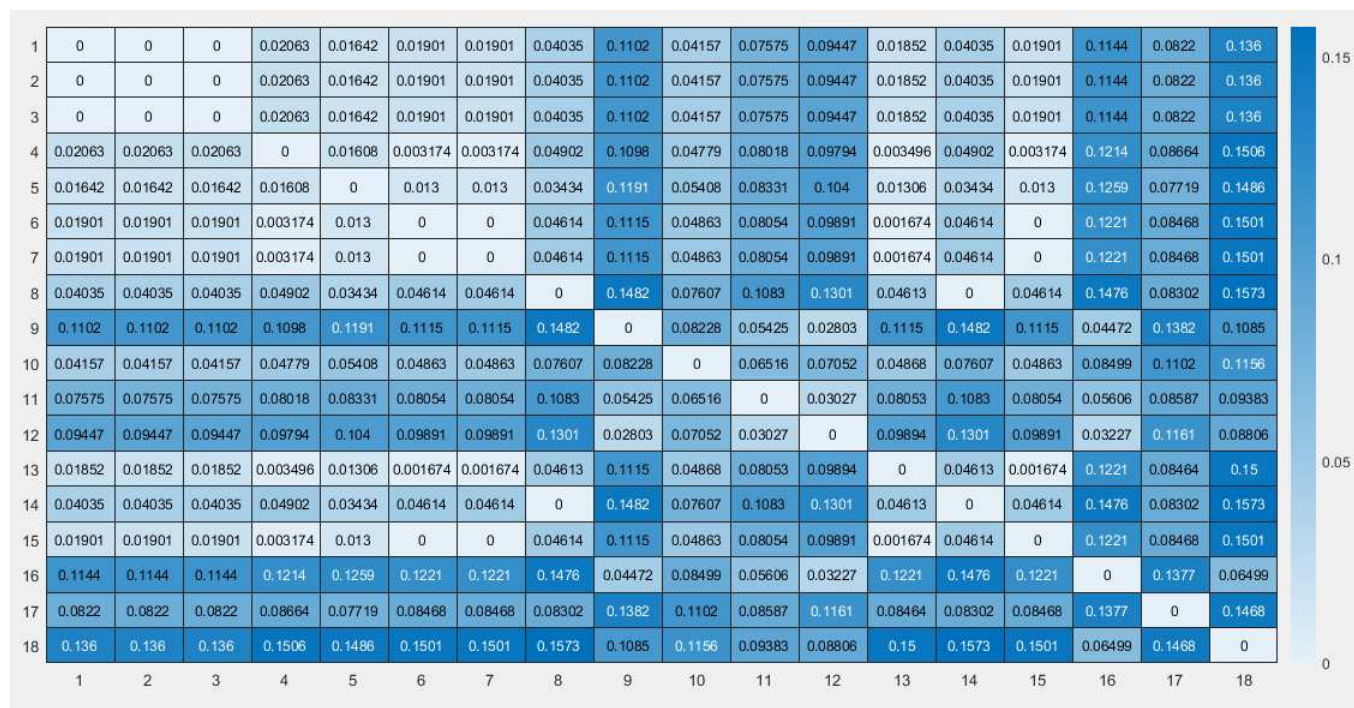Clusters based on bioinformatics of D2

Figure 9: Distance matrix based on the bioinformatics feature vectors of D2 of ACE2 across eighteen species and associated clusters
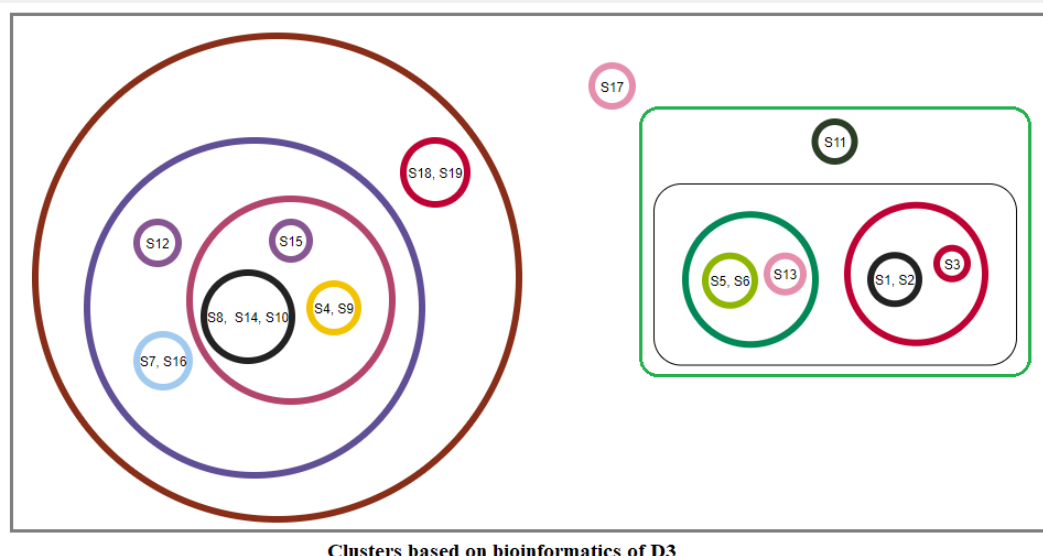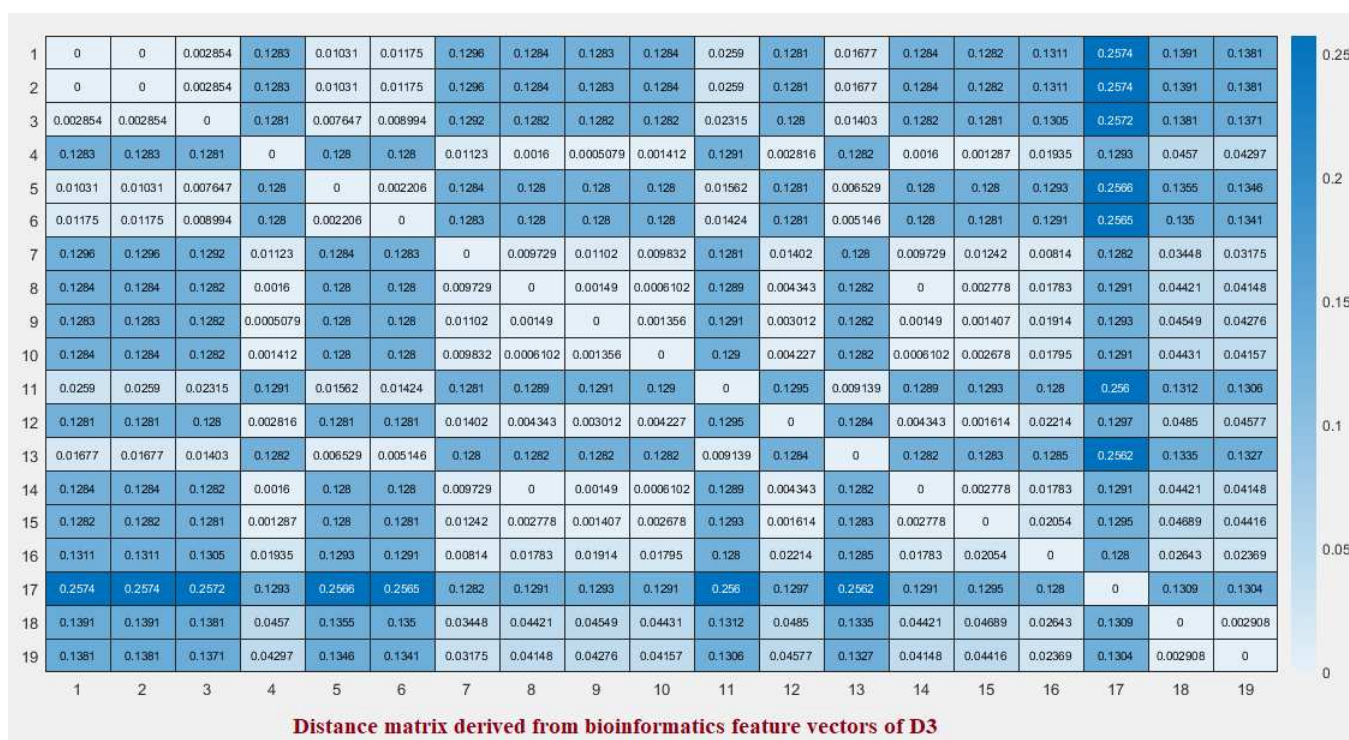
14

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	0	0.002854	0.1283	0.01031	0.01175	0.1296	0.1284	0.1283	0.1284	0.0259	0.1281	0.01677	0.1284	0.1282	0.1311	0.2574	0.1391	0.1381
2	0	0	0.002854	0.1283	0.01031	0.01175	0.1296	0.1284	0.1283	0.1284	0.0259	0.1281	0.01677	0.1284	0.1282	0.1311	0.2574	0.1391	0.1381
3	0.002854	0.002854	0	0.1281	0.007647	0.008994	0.1292	0.1282	0.1282	0.1282	0.02315	0.128	0.01403	0.1282	0.1281	0.1305	0.2572	0.1381	0.1371
4	0.1283	0.1283	0.1281	0	0.128	0.128	0.01123	0.0016	0.0005079	0.001412	0.1291	0.002816	0.1282	0.0016	0.001287	0.01935	0.1293	0.0457	0.04297
5	0.01031	0.01031	0.007647	0.128	0	0.002206	0.1284	0.128	0.128	0.128	0.01562	0.1281	0.006529	0.128	0.128	0.1293	0.2566	0.1355	0.1346
6	0.01175	0.01175	0.008994	0.128	0.002206	0	0.1283	0.128	0.128	0.128	0.01424	0.1281	0.005146	0.128	0.1281	0.1291	0.2565	0.135	0.1341
7	0.1296	0.1296	0.1292	0.01123	0.1284	0.1283	0	0.009729	0.01102	0.009832	0.1281	0.01402	0.128	0.009729	0.01242	0.00814	0.1282	0.03448	0.03175
8	0.1284	0.1284	0.1282	0.0016	0.128	0.128	0.009729	0	0.00149	0.0006102	0.1289	0.004343	0.1282	0	0.002778	0.01783	0.1291	0.04421	0.04148
9	0.1283	0.1283	0.1282	0.0005079	0.128	0.128	0.01102	0.00149	0	0.001356	0.1291	0.003012	0.1282	0.00149	0.001407	0.01914	0.1293	0.04549	0.04276
10	0.1284	0.1284	0.1282	0.001412	0.128	0.128	0.009832	0.0006102	0.001356	0	0.129	0.004227	0.1282	0.0006102	0.002678	0.01795	0.1291	0.04431	0.04157
11	0.0259	0.0259	0.02315	0.1291	0.01562	0.01424	0.1281	0.1289	0.1291	0.129	0	0.1295	0.009139	0.1289	0.1293	0.128	0.256	0.1312	0.1306
12	0.1281	0.1281	0.128	0.002816	0.1281	0.1281	0.01402	0.004343	0.003012	0.004227	0.1295	0	0.1284	0.004343	0.001614	0.02214	0.1297	0.0485	0.04577
13	0.01677	0.01677	0.01403	0.1282	0.006529	0.005146	0.128	0.1282	0.1282	0.1282	0.009139	0.1284	0	0.1282	0.1283	0.1285	0.2562	0.1335	0.1327
14	0.1284	0.1284	0.1282	0.0016	0.128	0.128	0.009729	0	0.00149	0.0006102	0.1289	0.004343	0.1282	0	0.002778	0.01783	0.1291	0.04421	0.04148
15	0.1282	0.1282	0.1281	0.001287	0.128	0.1281	0.01242	0.002778	0.001407	0.002678	0.1293	0.001614	0.1283	0.002778	0	0.02054	0.1295	0.04689	0.04416
16	0.1311	0.1311	0.1305	0.01935	0.1293	0.1291	0.00814	0.01783	0.01914	0.01795	0.128	0.02214	0.1285	0.01783	0.02054	0	0.128	0.02643	0.02369
17	0.2574	0.2574	0.2572	0.1293	0.2566	0.2565	0.1282	0.1291	0.1293	0.1291	0.256	0.1297	0.2562	0.1291	0.1295	0.128	0	0.1309	0.1304
18	0.1391	0.1391	0.1381	0.0457	0.1355	0.135	0.03448	0.04421	0.04549	0.04431	0.1312	0.0485	0.1335	0.04421	0.04689	0.02643	0.1309	0	0.002908
19	0.1381	0.1381	0.1371	0.04297	0.1346	0.1341	0.03175	0.04148	0.04276	0.04157	0.1306	0.04577	0.1327	0.04148	0.04416	0.02369	0.1304	0.002908	0

Distance matrix derived from bioinformatics feature vectors of D3

Clusters based on bioinformatics of D3

Figure 10: Distance matrix based on the bioinformatics feature vectors of D3 of ACE2 across nineteen species and associated clusters

A final set of six clusters was formed using the K-means clustering method to have all three domains for the nineteen different species (Fig. 11). Although the species S7 was clustered with the species S5 and S11 as per full-length ACE2 sequence homology, S7 formed a unique singleton cluster when the bioinformatics features were taken into consideration. Similarly, the species S16 formed a singleton cluster though it was clustered with S17, S18, and S19 as per amino acid homology of ACE2. The sequence homology of ACE2 made the four species S16, S17, S18, and S19 into a single cluster but bioinformatics features placed the species S18 in a cluster where the other three species S1, S2, and S3 belonged. Based on bioinformatics features, S4 clustered together with S15 though the ACE2 receptor of S4 was sequentially similar to ACE2 of S9, S10, and S12.
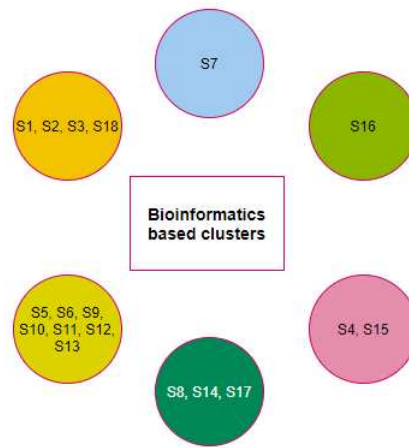
15

Figure 11: Clusters of species based on the bioinformatics of the D1, D2, and D3 domains

The clusters $\{S1, S2, S3\}$, $\{S6, S13\}$, and $\{S9, S10, S12\}$ were unaltered with respect to the full length ACE2 homology and bioinformatics features.

3.4. Phylogeny and clustering based on polarity

In the D1 domain, it was observed that across eighteen species, polarity of thirteen amino acids among nineteen (24-42 aa) amino acids were found to be conserved. Based on the polarity and non-polarity nature of the amino acids, the species were arranged in a phylogenetic tree (Fig. 12).

It was found that *Homo sapiens, Pan troglodyte, Macaca mulatta* and *Danio rerio* were closer according to this analysis. *Pteropus alecto, Pteropus vampyrus* and *Sus scrofa* were occurred in parallel along with the above three and formed a different clade indicating the closeness based on polarity. Again, the case for *Gallus gallus, Rhinolophus ferrumequinum, Mustela putorius furo, Equus caballus* and *Felis catus* is similar. Two separate groups, *Mesocricetus auratus, Manis javanica* and *Capra hircus, Bos taurus*, were similarly placed nearby, indicating that the polarity of amino acids of the proteins for these species was similar. *Pelodiscus sinensis* and *Rattus norvegicus* occurred separately and were not grouped with any other species but bears similarity with both the groups containing species *Mesocricetus auratus, Manis javanica and Pteropus alecto, Pteropus vampyrus, Sus scrofa.*

In the D2 domain, out of six amino acid long sequences, the polarity of three amino acids was conserved across 18 species, and among them, one amino acid was a binding residue. *Homo sapiens, Pan troglodytes, Macaca mulatta, Pteropus vampyrus* and *Pteropus alecto* were grouped together since the overall polarity of their amino acid chain was found to be similar and simultaneously, *Danio rerio, Mustela putorius furo, Gallus gallus* and *Pelodiscus sinensis* were placed together. Also, three groups comprising *Manis javanica, Capra hircus* & *Bos taurus, Sus scrofa* & *Rattus norvegicus, Rhinolophus ferrumequinum* respectively were placed in close proximity based on their polarity and non-polarity of the amino acids in the protein sequence. However, *Equus caballus, Felis cattus*, and *Mesocricetus auratus* were placed separately since they did not show much resemblance based on polarity.

16
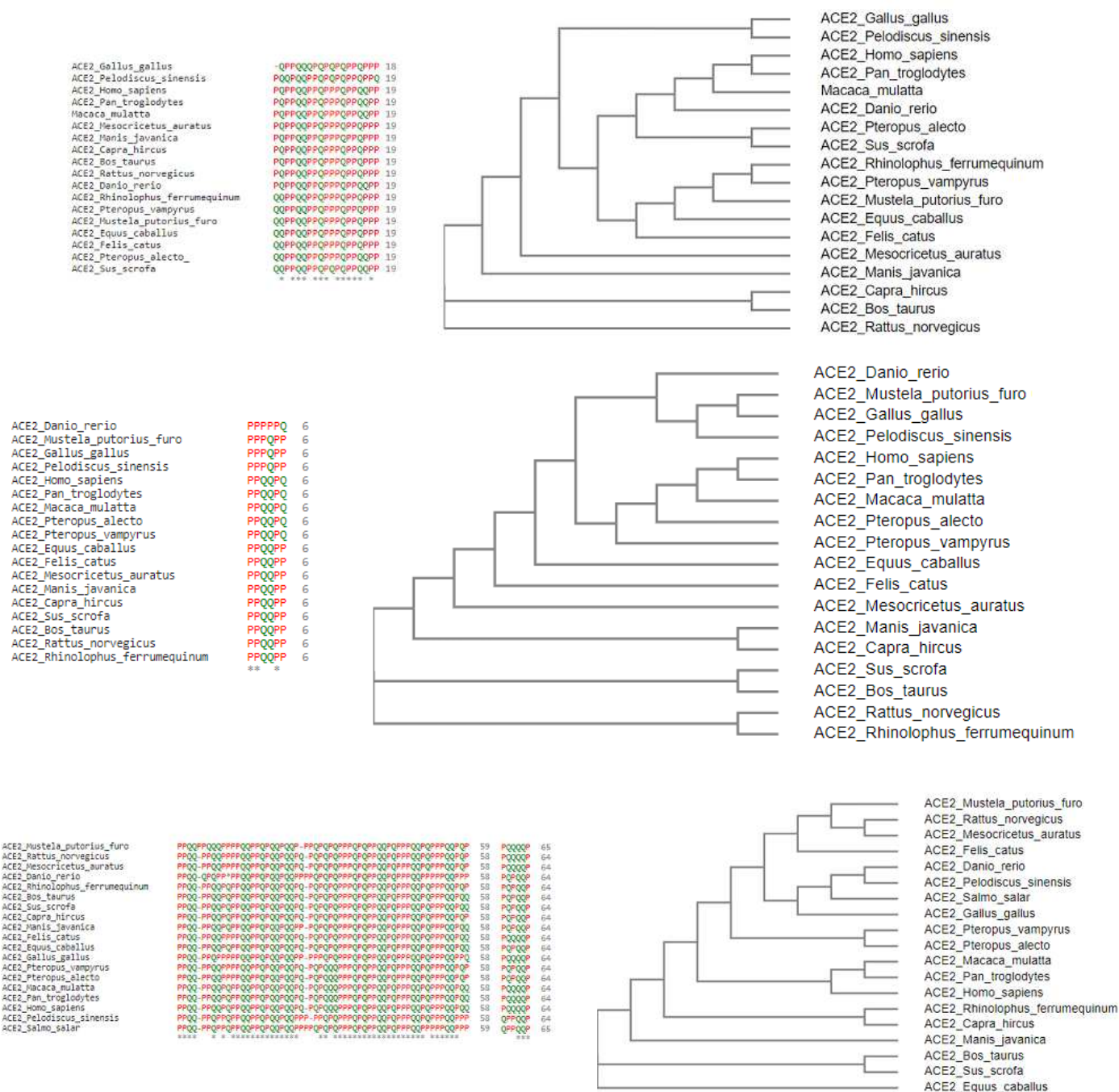
Figure 12: Polarity sequence of the D1, D2 and D3 domains across all species alignment and associated phylogenetic relationships

In the D3 domain of ACE2 sequences of *Salmo salar* and *Danio rerio* there was an insertion of a polar amino acid in one of the binding residue positions that may affect the binding of ACE2 to that of RBD of SARS-CoV2 negatively. A total of three binding residues were already reported in the D3 domain, of which one of them remained conserved concerning polarity across the nineteen species. *Rattus norvegicus, Mustela putorius furo, Mesocricetus auratus* and *Felis catus* were grouped under a single clade based on the polarity of their protein sequence. Similar was the case for *Danio rerio, Pelodiscus sinensis, Salmo salar*, and *Gallus gallus.* Due to the sequence similarity between *Pteropus vampyrus* and *Pteropus alecto*, their polarity of the protein sequence was also similar and thus grouped. Sequence similarity was also observed for *Homo sapiens, Pan troglodytes* and *Macaca mulatta*, so again these were categorized together. Two groups comprising of *Rhinolophus ferrumequinum, Capra hircus* and *Bos taurus, Sus scrofa* respectively, were sorted together indicating their similar nature of

polarity and non-polarity of protein sequence. Lastly, *Manis javanica* and *Equus caballus* were placed separately signifying that the sequence of both the species were quite distinct.

220   The individual groups of species based on the polarity of individual D1, D2 and D3 domains have emerged into six disjoint clusters (Fig. 13).
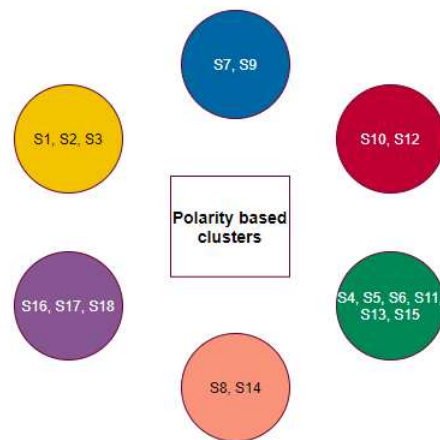
Figure 13: Clusters based on groups of species based on domain-wise polarity

Here the clusters $\{S1, S2, S3\}$, $\{S16, S17, S18\}$, $\{S8, S14\}$, $\{S6, S13\}$, and $\{S10, S12\}$ remained invariant with regards to the homology of full length ACE2 as well as polarity sequence of the D1, D2 and D3 domains.

3.5. Possible clusters of transmission of SARS-CoV-2

225   Based on all the different clustered formed on the basis of amino acid homology, secondary structures, bioinformatics, and polarity of the D1, D2, and D3 domains of ACE2, final clusters of all nineteen species were devised using the K-means clustering method Fig. 14.
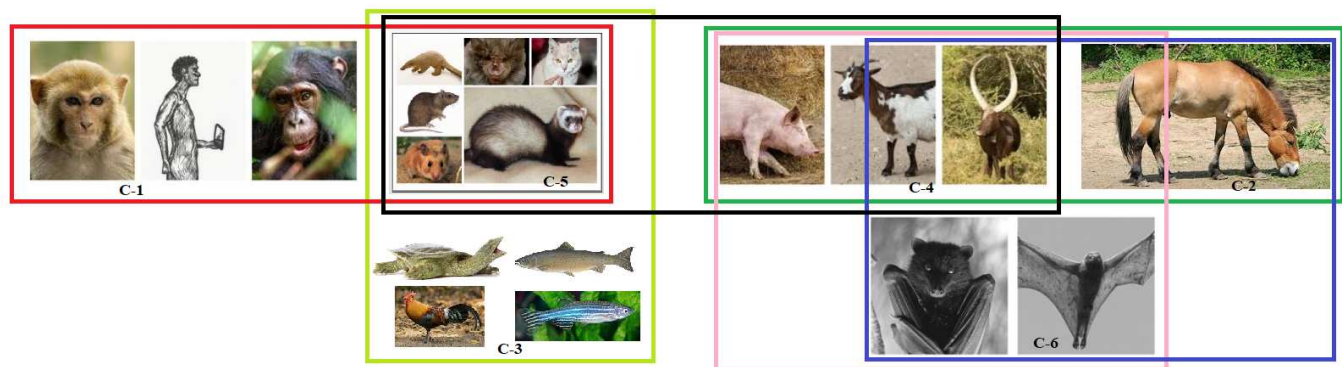
Figure 14: Schematic representation of a possible set of clusters of transmission of SARS-CoV-2

In Fig. 14, it was found that the cluster-1 (C-1) comprising of *Homo sapiens, Pan troglodyte*, and *Macaca mulatta* were close to cluster-5 (C-5) comprising of *Felis catus* (Cat), *Mesocricetus auratus* (Golden Hamster), *Manis javanica* (Sunda
230   pangolin), *Mustela putorius furo* (Ferret), *Rattus norvegicus* (Rat), and *Rhinolphus ferrumequinum* (Greater horseshoe bat). This C-5 is also close to cluster-3 (C-3) [*Gallus gallus* (red jungle fowl), *Pelodiscus sinensis* (Chinese shell turtle), *Danio rerio* (zebrafish) and *Salmo salar*], and cluster-4 (C-4) [*Capra hircus* (Goat), *Bos taurus* (Cattle) and *Sus scrofa* (pig)]. C-4

also showed resemblance with cluster-2 (C-2) [*Pteropus alecto*, and *Pteropus vampyrus*], and cluster-6 (C-6) that comprised

of *Equus caballus* (horse) only. Although, both C-2 and C-6 were also close to each other.

Furthermore, a pooled analyses based on the two type of substitutions (one is affecting SARS-CoV-2 transmission (M1), and the other one is SARS-CoV-2 non-affecting transmission(M2)) for all the six final clusters are presented in Table-3.

Table 3: M1 and M2 substitutions across nineteen ACE2 receptors

Cluster	Species	DI (M1, M2)	D2 (M1, M2)	D3 (M1, M2)	Total (M1, M2)
1	*Human ACE2*	0	0	0	0
1	*Pan troglodytes*	0	0	0	0
1	*Macaca mulatta*	0	0	0	0
6	*Equus caballus*	(4,1)	(1,0)	(0,0)	(5,1)
5	*Felis catus*	(1,2)	(1,0)	(0,0)	(2,2)
5	*Mesocrietus auratus*	(1,0)	(1,0)	(0,0)	(2,0)
5	*Manis javanica*	(2,2)	(1,1)	(1,0)	(4,3)
5	*Mustela putorius furo*	(2,2)	(2,0)	(1,0)	(5,2)
5	*Rattus norvegicus*	(3,1)	(3,1)	(0,1)	(6,3)
5	*Rhinolophus ferrumequinum*	(6,0)	(2,0)	(0,0)	(8,0)
4	*Capra hircus*	(0,1)	(2,0)	(0,0)	(2,1)
4	*Bos taurus*	(0,1)	(2,0)	(0,0)	(2,1)
4	*Sus scrofa*	(1,2)	(1,1)	(0,1)	(2,4)
2	*Pteropus vampyrus*	(2,1)	(1,0)	(1,1)	(4,2)
2	*Pteropus alecto*	(2,1)	(1,0)	(1,1)	(4,2)
3	*Gallus gallus*	(6,0)	(3,0)	(1,0)	(10,0)
3	*Pelodiscus sinensis*	(7,0)	(2,0)	(1,0)	(10,0)
3	*Danio rerio*	(5,0)	(2,0)	(2,1)	(9,1)
3	*Salmo salar*	NA	NA	(3,1)	(3,1)

Based on the Table-3 information regarding the number of M1 and M2 substitutions, the intra-species transmission of SARS-CoV-2 were enlightened as follow:

- C-1: None of the species bear any mutation in the binding residues and are conserved, so viral transmission is immaculate.

- C-2: This cluster has an equal number of transmission affecting and transmission non-affecting types of substitutions. Therefore, both have an equal probability of getting infected from each other.

- C-3: Here again, *Gallus gallus, Pelodiscus sinensis*, and *Danio rerio* have a similar ratio of S1 to S2, signifying possible flow of viral transmission within these three species. However, *Salmo salar* is unique and distant, and therefore, the probability of viral transmission is unlikely.

- C-4: The species in this cluster have a similar number of transmission-affecting and transmission non-affecting type of substitutions which shows that the flow of viral transmission would be continuous among these three species.

- C-5: Transmission between *Felis catus* and *Mesocrietus auratus* is highly likely which is the same for *Manis javanica, Mustela putorius furo*, and *Rattus norvegicus* as indicated by their similar number of substitutions. Therefore, the inter-transmission between these species is highly plausible. While *Rhinolophus ferrumequinum* has a relatively high value of transmission affecting substitutions from all of the above, its susceptibility to getting infected from others species is uncertain.

- C-6: A total of five transmission affecting substitutions in the three domains for human were observed.

4. Discussion

In this study, we amassed the ACE2 protein sequences of nineteen species to investigate the possible transmission of SARS-CoV-2 among these species in relation to human ACE2 protein. Multiple sequence alignments of these ACE2 receptors

19

enabled us to estimate the similarity concerning amino acids and from that, we observed that *Salmo salar* (Salmon fish) was quite distant. It also gave us the idea that some of the amino acid substitutions in the binding residues occurring across the species with respect to human ACE2 were resulted in amino acids having similar binding properties, indicating that their interactions with RBD of the S protein will be similar to that of humans, thus making transmission across these species feasible. It was observed that *Homo sapiens* and *Pan troglodytes* (Chimpanzee) have complete sequence similarity. In contrast, *Macaca mulatta* (Rhesus macaque) shared a high percentage of sequence identity except for two amino acid positions. However, no substitutions were observed in the binding of amino acid residues, making the viral transmission across these species highly likely. Again, *Pteropus vampyrus* (Large flying fox) and *Pteropus alecto* (Black flying fox) have precisely the same ACE2 sequence and thus signifying high viral transmission and that both of them have an equal chance of getting infected with each other.

Further analysis led us to present a possible transmission flow among the nineteen species, as illustrated in Fig. 14. The multifaceted examination of the ACE2 protein indicated that interspecies SARS-CoV-2 transmission is quite possible and we have tried to provide a better insight into it by predicting the possible transmission among species within the same cluster and between clusters too. However, further in-depth analysis is necessary in the future for the identification of new hosts of SARS-CoV-2 as well as for determination of possible ways to prevent inter-species transmission.

Author Contributions

SSH conceived the problem. SSH, DA, SG carried out the work. SSH analyzed the results and wrote the primary draft of the article. All authors reviewed, edited, and approved the final manuscript.

Conflict of Interests

The authors do not have any conflicts of interest to declare.

References

[1] W. H. Organization, Severe acute respiratory syndrome (sars), Weekly Epidemiological Record 78 (13) (2003) 89–89.

[2] W. H. Organization, Outbreak news: severe acute respiratory syndrome (sars), Weekly Epidemiological Record 78 (12) (2003) 81–83.

[3] J.-j. Zhang, X. Dong, Y.-y. Cao, Y.-d. Yuan, Y.-b. Yang, Y.-q. Yan, C. A. Akdis, Y.-d. Gao, Clinical characteristics of 140 patients infected with sars-cov-2 in wuhan, china, Allergy (2020).

[4] Y. Liu, Z. Ning, Y. Chen, M. Guo, Y. Liu, N. K. Gali, L. Sun, Y. Duan, J. Cai, D. Westerdahl, et al., Aerodynamic analysis of sars-cov-2 in two wuhan hospitals, Nature 582 (7813) (2020) 557–560.

[5] W. H. Organization, et al., Coronavirus disease 2019 (covid-19): situation report, 82 (2020).

[6] Y.-Z. Zhang, E. C. Holmes, A genomic perspective on the origin and emergence of sars-cov-2, Cell (2020).

[7] M. Zheng, L. Song, Novel antibody epitopes dominate the antigenicity of spike glycoprotein in sars-cov-2 compared to sars-cov, Cellular & molecular immunology 17 (5) (2020) 536–538.

[8] R. Minakshi, A. T. Jan, S. Rahman, J. Kim, A testimony of the surgent sars-cov-2 in the immunological panorama of the human host, Frontiers in Cellular and Infection Microbiology 10 (2020) 539.

[9] X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, et al., Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study, The Lancet Respiratory Medicine (2020).

[10] S. Zaim, J. H. Chong, V. Sankaranarayanan, A. Harky, Covid-19 and multi-organ response, Current Problems in Cardiology (2020) 100618.

[11] F. Saponaro, G. Rutigliano, S. Sestito, L. Bandini, B. Storti, R. Bizzarri, R. Zucchi, Ace2 in the era of sars-cov-2: controversies and novel perspectives (2020).

[12] S. Li, J. Han, A. Zhang, Y. Han, M. Chen, Z. Liu, M. Shao, W. Cao, Exploring the demographics and clinical characteristics related to the expression of angiotensin-converting enzyme 2, a receptor of sars-cov-2, Frontiers in Medicine 7 (2020) 530.

[13] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, et al., Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor, Nature 581 (7807) (2020) 215–220.

[14] M. Hussain, N. Jabeen, F. Raza, S. Shabbir, A. A. Baig, A. Amanullah, B. Aziz, Structural variations in human ace2 may influence its binding with sars-cov-2 spike protein, Journal of medical virology (2020).

[15] P. McMillan, B. D. Uhal, Covid-19–a theory of autoimmunity to ace-2, MOJ immunology 7 (1) (2020) 17.

[16] L. Samavati, B. D. Uhal, Ace2, much more than just a receptor for sars-cov-2, Frontiers in Cellular and Infection Microbiology 10 (2020) 317.

[17] G. K. Veeramachaneni, V. Thunuguntla, J. Bobbillapati, J. S. Bondili, Structural and simulation analysis of hotspot residues interactions of sars-cov 2 with human ace2 receptor, Journal of Biomolecular Structure and Dynamics (2020) 1–11.

[18] W. Li, C. Zhang, J. Sui, J. H. Kuhn, M. J. Moore, S. Luo, S.-K. Wong, I.-C. Huang, K. Xu, N. Vasilieva, et al., Receptor and viral determinants of sars-coronavirus adaptation to human ace2, The EMBO journal 24 (8) (2005) 1634–1643.

[19] M. Gheblawi, K. Wang, A. Viveiros, Q. Nguyen, J.-C. Zhong, A. J. Turner, M. K. Raizada, M. B. Grant, G. Y. Oudit, Angiotensin-converting enzyme 2: Sars-cov-2 receptor and regulator of the renin-angiotensin system: celebrating the 20th anniversary of the discovery of ace2, Circulation research 126 (10) (2020) 1456–1474.

[20] C. Tikellis, M. Thomas, Angiotensin-converting enzyme 2 (ace2) is a key modulator of the renin angiotensin system in health and disease, International journal of peptides 2012 (2012).

[21] K. Gorshkov, K. Susumu, J. Chen, M. Xu, M. Pradhan, W. Zhu, X. Hu, J. C. Breger, M. Wolak, E. Oh, Quantum dot-conjugated sars-cov-2 spike pseudo-virions enable tracking of angiotensin converting enzyme 2 binding and endocytosis, ACS nano (2020).

[22] V. N. Uversky, F. Elrashdy, A. Aljadawi, E. M. Redwan, Household pets and sars-cov2 transmissibility in the light of the ace2 intrinsic disorder status, Journal of Biomolecular Structure and Dynamics (2020) 1–4.

[23] K. D. Pruitt, T. Tatusova, D. R. Maglott, Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic acids research 33 (suppl_1) (2005) D501–D504.

[24] S. Federhen, The ncbi taxonomy database, Nucleic acids research 40 (D1) (2012) D136–D143.

[25] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T. L. Madden, Ncbi blast: a better web interface, Nucleic acids research 36 (suppl_2) (2008) W5–W9.

[26] J. P. Jenuth, The ncbi, in: Bioinformatics Methods and Protocols, Springer, 2000, pp. 301–312.

[27] A. Likas, N. Vlassis, J. J. Verbeek, The global k-means clustering algorithm, Pattern recognition 36 (2) (2003) 451–461.

[28] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, S. J. Brown, Fgka: A fast genetic k-means clustering algorithm, in: Proceedings of the 2004 ACM symposium on Applied computing, 2004, pp. 622–623.

[29] T. A. Kumar, Cfssp: Chou and fasman secondary structure prediction server, Wide Spectrum 1 (9) (2013) 15–19.

[30] A. Pande, S. Patiyal, A. Lathwal, C. Arora, D. Kaur, A. Dhall, G. Mishra, H. Kaur, N. Sharma, S. Jain, et al., Computing wide range of protein/peptide features from their sequence and structure, bioRxiv (2019) 599126.

[31] V. K. Garg, H. Avashthi, A. Tiwari, P. A. Jain, P. W. Ramkete, A. M. Kayastha, V. K. Singh, Mfppi–multi fasta protparam interface, Bioinformation 12 (2) (2016) 74.

[32] C. N. Pace, F. Vajdos, L. Fee, G. Grimsley, T. Gray, How to measure and predict the molar absorption coefficient of a protein, Protein science 4 (11) (1995) 2411–2423.

[33] H.-X. Zhou, X. Pang, Electrostatic interactions in protein structure, folding, binding, and condensation, Chemical reviews 118 (4) (2018) 1691–1741.

[34] M. M. Flocco, S. L. Mowbray, Strange bedfellows: interactions between acidic side-chains in proteins, Journal of molecular biology 254 (1) (1995) 96–105.